Characterization of time-varying regimes in remote sensing time series: application to the forecasting of satellite-derived suspended matter concentrations

Bertrand Saulquin, Ronan Fablet, Pierre Ailliot, Grégoire Mercier, David Doxaran, Antoine Mangin, Odile Fanton d'Andon

Characterization of time-varying regimes in remote sensing time series: application to the forecasting of satellite-derived suspended matter concentrations

Bertrand Saulquin (1, 2), Ronan Fablet (2, 3), Pierre Ailliot (3), Grégoire Mercier (2, 3), David Doxaran (4), Antoine Mangin (2), Odile Fanton d'Andon (2).

(1) ACRI-ST, Sophia-Antipolis, 260 route du Pin Montard, BP 234 06904 Sophia-Antipolis, France

(2) Institut Mines-Telecom, Télécom Bretagne; UMR CNRS 3192 Lab-STICC, Technopôle Brest Iroise CS 83818, 29238 Brest, France

(3) Université Européenne de Bretagne, 35000 Rennes, France

(4) CNRS, Université Pierre et Marie Curie-Paris 6, UMR 7093, Laboratoire d'Océanographie de Villefranche/Mer, 06230 Villefranche-sur-Mer, France

Corresponding author: bertrand.saulquin@acri-st.fr

Submission date: 19/05/2014

Abstract

Satellite data, with their spatial and temporal coverage, are particularly well suited for the analysis and characterization of space-time-varying relationships between geophysical processes. In this study, we investigate the forecasting of a geophysical variable using both satellite observations and model outputs. The studied latent-regime models aim here at identifying time-varying regime

shifts within a dataset which is a key of interest for geophysical processes driven by the seasonal variability. As a specific example, we study the daily concentration from 2007 to 2009 of mineral suspended particulate matters estimated from the satellite-derived MODIS, MERIS and SeaWiFS dataset, in coastal waters adjacent to the Gironde River mouth (South West of France). We clearly show that the forecast of the high resolution suspended particulate matter dataset using environmental data (wave height, wind strength and direction, tides and river outflow) and a multi-regime model is significantly improved compared with a classical multi-regression and a Support Vector Regression model. Each regime is here characterized by a regression function and a covariance structure.

From an analytical point of view, we compare the results obtained with four models: homogeneous and non-homogeneous Markov-switching models, with and without an autoregressive term, i.e. the suspended matter concentration observed the day before. Inclusion of an autoregressive term is motivated by the strong natural autocorrelation level depicted by geophysical time series, but, one may avoid this term if, for example, the observations are no more available during specific conditions or periods. With the evaluated models, best results are obtained with a mixture of 3 regimes for both autoregressive and non-autoregressive models. Prediction performance at day+1, using the non-autoregressive models and a validation dataset, reached 80% of the observed variance, compared to 32% for a standard single-regime (regression) analysis, and 40 % for a Support Vector Regression. Inclusion of an autoregressive term increases results to 93% of explained variance for the mixture model compared to 80% without autoregressive term and 85% using a Support Vector Regression. These results stress the potential of the identification of geophysical regimes to improve the forecasting, or the inversion, of a high resolution geophysical variable using both observations and model outputs. We also show that for short periods of lack of observations (less than 15 days), estimations using the autoregressive term are better than without. In this case the autoregressive term and the transition probabilities between regimes are estimated using available model outputs.

Index term: 1) Satellite-derived suspended matter time series analysis. 2) Statistical forecasting. 3) Regime-switching latent regression models. 4) Joint analysis of satellite-derived products and operational model outputs. 5) Gironde river plume.

1 Introduction

The forecasting of a geophysical variable using statistical models is an alternative to model-based approaches which typically involve complex simulation and/or assimilation [1, 2]. For instance, coupled hydrodynamic and sediment transport models can be used to estimate the concentration of suspended particulate matters within the water column [3] while statistical approaches may use available satellite and model data to predict the same variable [4]. Many statistical approaches have been proposed and evaluated to forecast or infer a studied variable from predictors. Among them, linear multivariate regression [5] and non-linear (polynomial) multivariate regression [6] are the most known. Numerous specific models dedicated to time-series analysis such as AutoRegressive Moving Average (ARMA) and AutoRegressive Integrated Moving Average (ARIMA) models [7] have also been developed initially to address financial time series. These latest, which aim at studying the behavior of a time series without considering forcing factors, have also been applied to geophysical time series [8]. Non-linear regressions, based on supervised learning strategies, such as Neural Networks [9] and Support Vector Regressions (SVR) [10] may provide relevant alternatives to estimate a variable from predictors. In the context of geophysical studies, they may nevertheless suffer from two major drawbacks. First, though relevant regression performances may be reported, these models are not physically interpretable and may be very sensitive to the training dataset. Second, multi-regime dynamics, often exhibited by geophysical processes driven by the seasonality [11], cannot be addressed by non-linear models, contrary to latent-regime models as demonstrated in our study.

We propose here to characterize time-varying relationships between a variable and its forcing parameters using latent-regime models, and hence optimize forecasting results. As an illustration, we address the concentration of inorganic suspended particle matters (SPIM), estimated from satellite data using a regional algorithm [12, 13], and observed in the mouth of the Gironde estuary. In this area, sediments are mainly exported from the Gironde estuary [13, 14] and SPIM concentration clearly depends on the local physical forcing: swell, tide, wind and river discharge. A minimum of energy has to be brought by waves and tides to re-suspend cohesive sediments accumulated at the bottom. Conversely, when sediments have been re-suspended in the water column by wave influence, their settling velocity depends on their size and density [15] and physico-chemical properties [16]. This example stresses that the relationships between the studied variable (SPIM) and the causing factors evolve in space and time and potentially requires advanced statistical methods to identify the underlying geophysical regimes.

From a methodological point of view, "latent regime regressions" also referred as "clusterwise regressions" [17, 18] are particularly appealing to identify such non-linear and multi-regime patterns within a dataset. Each regime is associated with a linear regression and a non-linear relationship is thus estimated as a sum of linear contributions. Regarding the temporal dynamics of these regimes, we here consider Markovian processes [19], which state the transitions in time between two regimes. The standard Hidden Markov Model (HMM) and Non-Homogeneous Markov Model (NHMM) are evaluated [19]. The inclusion of an autoregressive term (HMM-AR) and (NHMM-AR) is also discussed. This aspect is motivated on the one hand by the strong autocorrelation level depicted by geophysical time series [20]. When the observation at t-1 is available, it is obvious, considering the strong natural autocorrelation of geophysical data that the forecast at time t should take into account the observation at time t-1. Conversely, for specific applications, or if the observations are not available during long periods (such as winter storms, or after a sensor failure), one may need to estimate the variable without using the observations of the

previous days. We discuss here the choice between autoregressive or not autoregressive models for long lack of observation period using forecasting results from t+1 to t+15.

Model parameter estimation is carried out from a dataset composed of 5862 time series of 1096 points in the mouth of the Gironde estuary in the [3°W-1°E ; 45-46.5°N] area during the period 2007-2009. Validation is performed on the same area for using the data for the year 2010. We used EOFs to reduce the dimension of the space-time observations. This is a usual approach in spatio-temporal statistics [21, 22] although alternatives may be considered such as linear discriminant analysis [23], and, we could also introduce a latent variable to describe the regime at each location and interact with the regimes at other locations. Nevertheless, such models are known to be very difficult to fit on the data and remain a research challenge for statisticians. We infer our mixture model using the expansion coefficients of the first four modes of the EOF which explain 99% of the total variance. The variables used as predictors for the SPIM expansion coefficients (EC) are the wave height issued from a numerical model [24], the wind fields optimally interpolated from satellite observations [25], the tide coefficient [26] and the Gironde fresh water discharge (sum of the Garonne and Dordogne Rivers contributions).

2 Methods

2.1 Markov switching forecast models

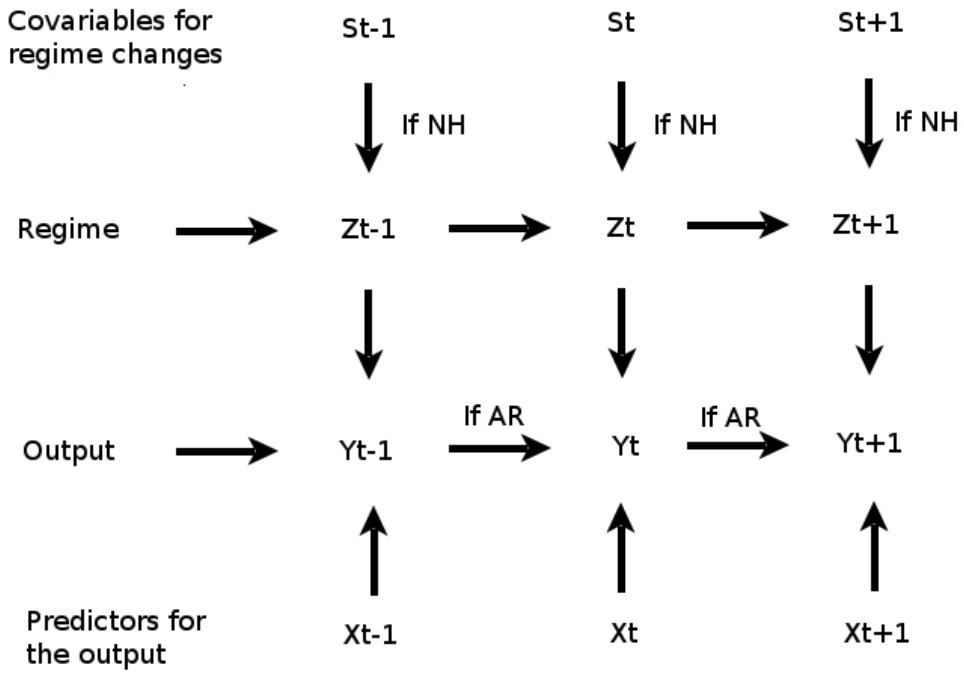
We address here the study of a two dimensional scalar geophysical time series Y. In a hidden Markov model framework (HMM; [19]), one states two different processes, the observed process Y and a hidden process Z. The observed process (here the turbidity) is assumed to be temporally dependent of the hidden process. The hidden process Z_t is modeled as a first order Markov chain [19]. At a given time t, the hidden variable $Z_t = k$ is a discrete value which states the regime

119 characterized by a latent [17] regression model with coefficient B_k between the variable Y_t and the
120 predictor X_t. At time t, knowing regime variable Z_t, the observed variable Y_t is modeled as:

121 $$(Yt \mid Zt=k) = XtBk \quad (1)$$

122 where $XtBk$ is the regression function, which predicts variable Y_t from some predictors X_t for
123 regime $Z_t = k$.

124 Figure 1 shows a graphical representation of the conditional dependencies involved in the
125 model, in the form of the general Directed Acyclic Diagram (DAG). It illustrates the interactions
126 between the variable $Y_{t,}$, the predictors $X_{t,}$ the hidden regime Z_t and the covariate S_t which acts on
127 regime switching. X_t may contain lagged values of Y_t (referred as autoregressive terms) and/or
128 exogenous covariates such as wind or wave height. Figure 1 defines a general family of model
129 which encompasses the most usual ones with regime switching. When no covariate is considered
130 i.e., Z_t only depends $Z_{t-1,}$ and, $Y_{t,}$ only depends on ($Y_{t,-s}$ ..$Y_{t,-1}$) and Z_t, we retrieve the usual Markov
131 switching autoregressive (MS-AR). If we further assume that s=0 (without autoregressive
132 component where $Y_{t,}$ only depends on Z_t) then we obtain the Hidden Markov Models (HMMs).
133 When Z_t does not dependent on $Z_{t-1,}$ and the dependence on S_t is parameterized using indicator
134 functions, we obtain the threshold autoregressive (TAR) model which is the other important family
135 with regime switches in the literature. HMMs, MS-AR and TAR have been used in many fields of
136 applications including geosciences [27].



137

138 Figure 1: Graphical representation of the various Markov-Switching Models considered in this
139 work: the arrows state the conditional dependencies between the random processes in play, namely
140 hidden regime process Z, observed process Y, prediction process X and regime change covariate
141 process S.

142 In the following equations (2 to 14), X_t and S_t are known, as they are either observations or
143 numerical model outputs. Given (1) the conditional likelihood of the observation Y_t given
144 predictors X_t and regime Z_t is expressed as [17]:

145 $$PYtXt,Zt=k\sim N(XtBk,\sigma k) \quad (2)$$

146 where N represents the Gaussian probability density function with mean XtBk and covariance
147 σ_k. Hence, given predictors up to time t we can predict process Y from its expectation conditionally
148 to process X:

149 $Yt=E\,YtXt=k=1KPZt=kXt\,.$ EYtXt,Zt=k=$k=1KPZt=kXt\,.$ XtBk (3)

150 where K is the number of regimes. $PZt=kXt=\pi tk$ is the posterior probability that the
151 dynamical regime Z at time t is of type k [17].

152 *3 Markov-switching priors*

153

154 Stating hidden regime process Z as a first-order Markovian process amounts to modeling the
155 transition between successive regimes at time t and time t-1. In the simplest case, one assumes
156 homogeneous transitions, i.e. time-independent and context-independent transitions, and the
157 Markovian process is fully characterized by its transition matrix $P(Z_t=k \mid Z_{t-1}=l)$ for possible pairs
158 of successive regimes (k, l). In the HMM setting the conditional distribution of Z_t, given the past
159 values Y_s and Z_s for s<t, is assumed to depend only on Z_{t-1} (Fig. 1):

160 $\pi tk=PZt=k\;Y0..Yt)=\;l=1KPYt\;\;Zt=k.P(Zt=kZt-1=l.PZt-1=l\;Y0..Yt-1)$ (4)

161 A NHMM extends this idea by allowing the transition matrix between the hidden states to
162 depend on a set of observed covariates S_t. Hughes and Guttorp [28, 29] highlighted the added value
163 of the NHMM to characterize the links between the large-scale atmospheric measures and the
164 small-scale spatially discontinuous precipitation field. In the NHMM settings, the transition matrix
165 between states $P(Zt=kZt-1=l$ in (3) is now time-dependent and conditioned by the covariates S_t:

166 $\pi tk=PZt=k\;Y0,..,Yt,\;St)=$

167 $l=1KPYt\;\;Zt=k,.P(Zt=kZt-1=l,\;St.PZt-1=l\;\;Y0\;..Yt-1)$ (5)

168 with:

169 $P(Zt=k\ Zt-1=l,\ \mathrm{St}=P\mathrm{St}\ Zt,=k,\ Zt-1=l\ .\ PZt=kZt-1=l\ /$

170 $k,lP\mathrm{St}\ Zt=k\ ,\ Zt-1=l\ .\ PZt=kZt-1=l$ (6)

171 The non-homogeneous transition between states is derived from the likelihood of the covariate
172 S_t given the state transition (Z_t, Z_{t+1}). We suppose that the probability density function of the
173 covariates during this change of state follows a normal distribution:

174 $\mathrm{PSt}\ \mathrm{Zt}=k,\ \mathrm{Zt}-1=l=N\mu l, k, \Sigma l, k$ (7)

175 Where N is a multivariate normal distribution with n means $\theta s=\mu l, k, \Sigma l, k$, μ is here of
176 dimension n, the number of covariates used to estimate the transitions. For a 'standard' multivariate
177 gaussian distribution $\Sigma_{l,k}$ is a covariance matrix. In the present application, and to reduce the
178 number of parameters to be estimated, we consider that the predictors are uncorrelated (null
179 covariance) and their relative influence is identical (same variance), i.e. $\Sigma_{l,k}$ is a multiple of the
180 identity matrix.

181

182 ## *3.1 Estimation of the model parameters*

183

184 The considered models involve two categories of parameters: those of the observation model, $\theta_{k,}$
185 namely regression coefficient B_k and standard deviation σ_k for each regime (Eq. 2) and those of the
186 Markov-swtiching prior, namely θ_s (Eq. 5). Given observed Y and X series, we proceed to the
187 estimation of model parameters according to a classical maximum likelihood (ML) criterion using
188 an iterative Expectation Maximisation (EM) procedure [30] expressed here without covariates:

189 $L\theta=t\mathrm{PYtY0}..\ \mathrm{Yt}-1, \theta$ (8)

where $\theta = \{\theta_s, \theta_k\}$ is the set of parameters to be estimated. For a given initialization for the parameters the EM procedure iterates estimation steps (E-step) of the posterior regime likelihood πtk with the given modes and the maximisation step (M-step) for the update of the parameters given these posteriors. The algorithm iterates until convergence between steps n and n+1, i.e. $|L\theta(n)-L\theta(n+1)|<10-3$. The posterior likelihood πtk (Eq 3&4) of the latent regime Z_t, is estimated in the E-step using the classical forward-backward recursions [31, 32] given series X and Y and current parameter estimate $\theta_k^{(n)}$, $\theta_s^{(n)}$. The M-step re-estimates the parameters $\theta_k^{(n+1)}$, $\theta_s^{(k+1)}$. For this, it is often possible to break the optimization problem into several lower dimensional optimization problems which are much quicker to solve [32]. More precisely, for all the models considered in this paper, it is possible to separate the parameters related to the evolution of the hidden Markov chain θ_s, and the parameters related to the evolution of the observed process in each regime θ_k:

$$\theta = argmax\theta s \; t\log(P(Zt=k\; Zt-1=l,\; St,\theta s(n))PZt=k,Zt-1=l\; Y0..\; YT,\; St,\; \theta(n) \qquad (9)$$

$$\theta k \;\; Bk \;\; (n+1)=argmin Bkt PZt=k Y0..\; YT,\theta(n))\; (Yt-BkXt2) \qquad (10)$$

$$\sigma k \;\; (n+1)=t PZt=k Y0..\; YT,\theta(n))\; (Yt-BkXt2) \qquad (11)$$

3.2 Forecasting application

The considered multi-regime regression models are applied to the short-term forecasting of series Y. More precisely, at a given time t, we aim at predicting variable Y at time t+dt. We typically assume that prediction variables X and covariates S, typically numerical simulations, are available up to time t+dt whereas the variable Y is only known up to time t. is estimated using $X_{0..t+dt}$ and $S_{0..\ +dt}$ (for inhomogeneous transitions). Thus, Y*t*+dt is given by the conditional

expectation of variable Y_{t+dt} given observations series up to time t and predictor series up to time t+ dt:

(12)

For HMM and NHMM it resorts to:

(13)

For autoregressive models HMM-AR and NHMM-AR, i.e. X_{t+dt} contains Y_{t+dt-1} which is not available, is estimated using $Y_{t..}$, X_t .. X_{t+dt-1} and . Estimated Y*t*+dt resorts to:

(14)

It might be noted that these predictions actually account for the uncertainties in the determination of the underlying regimes. Contrary to deterministic methods, confidence interval and uncertainties on Y*t*+dt can be derived [33] which is a key issue for modeling considerations.

3.3 Model performance estimation

A key issue in practice, which has received lots of attention in the last few years, is the problem of model selection which aims at finding the ”optimal” number of predictors and covariates [31]. Hereafter, we have chosen to use both the Bayes Information Criterion (BIC) and the explained variance (EVAR) as a first guides. BIC index generally permits to select parsimonious models which fit the data well [34]. It is defined as:

$$BIC = -2\log^*(L) + p^*\log(S) \quad (15)$$

Where L is the likelihood of the data, p is the number of parameters and S is the number of observations. The likelihood which is an output of the backward-forward recursions performed in the E-step. We also use the classical explained variance, EVAR, to characterize the model relevance:

$$\mathrm{EVAR}=1.-\mathrm{var}\,Y_{t+1}-Y_{t+1}\,/\,\mathrm{var}(Y_{t+1}) \quad (16)$$

BIC and EVAR are partially linked [34]. BIC tends to penalize complex models whereas explained variance criterion only qualifies the result and may lead to the over-parameterization of a model that typically lead to errors when other dataset are tested using the same parameterization. Therefore we consider both BIC and EVAR to assess the model performance.

The choice of the predictors and the covariates is performed here as follows. We first select as predictors the variable showing a significant correlation with the studied variable. Given these predictor datasets, we tested all the possible configurations and chose the predictors which provide the lower BIC on the training dataset and the greatest EVAR using the training (EVAR_train) and the validation dataset (EVAR_valid).

4 The data

4.1 The studied variable

Non-algal SPM concentrations (SPIM) are estimated using an analytical algorithm [12] defined as the difference between total SPM and phytoplankton biomass, the latter derived from Chl-a. It incorporates mainly mineral SPM and smaller amounts of organic SPM not related to living

255 phytoplankton. This method to derive non-algal SPM from remote-sensing reflectance is based on
256 the inversion of a simplified equation of radiative transfer, assuming that chlorophyll concentration
257 is known. This merged dataset consists of fields of non-algal surface SPM concentrations, derived
258 from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS), the Moderate Resolution Imaging
259 Spectroradiometer (MODIS) and the Medium Resolution Imaging Spectrometer (MERIS) sensors,
260 provided by the Ocean Colour TAC (Thematic Application Facility) of MyOcean, and interpolated
261 with a kriging method [35] for the period 2007–2009 over the Gironde mouth river from 3°W-1°E ;
262 45-46.5°N. Finally 5682 continuous time series of 1096 days compose our initial dataset of mineral
263 suspended matters concentration.

264 We first account for the space-time variability of the dataset, previously detrended and centered
265 for each time series [37] using a EOF decomposition [21], expressed here using the matrix form:

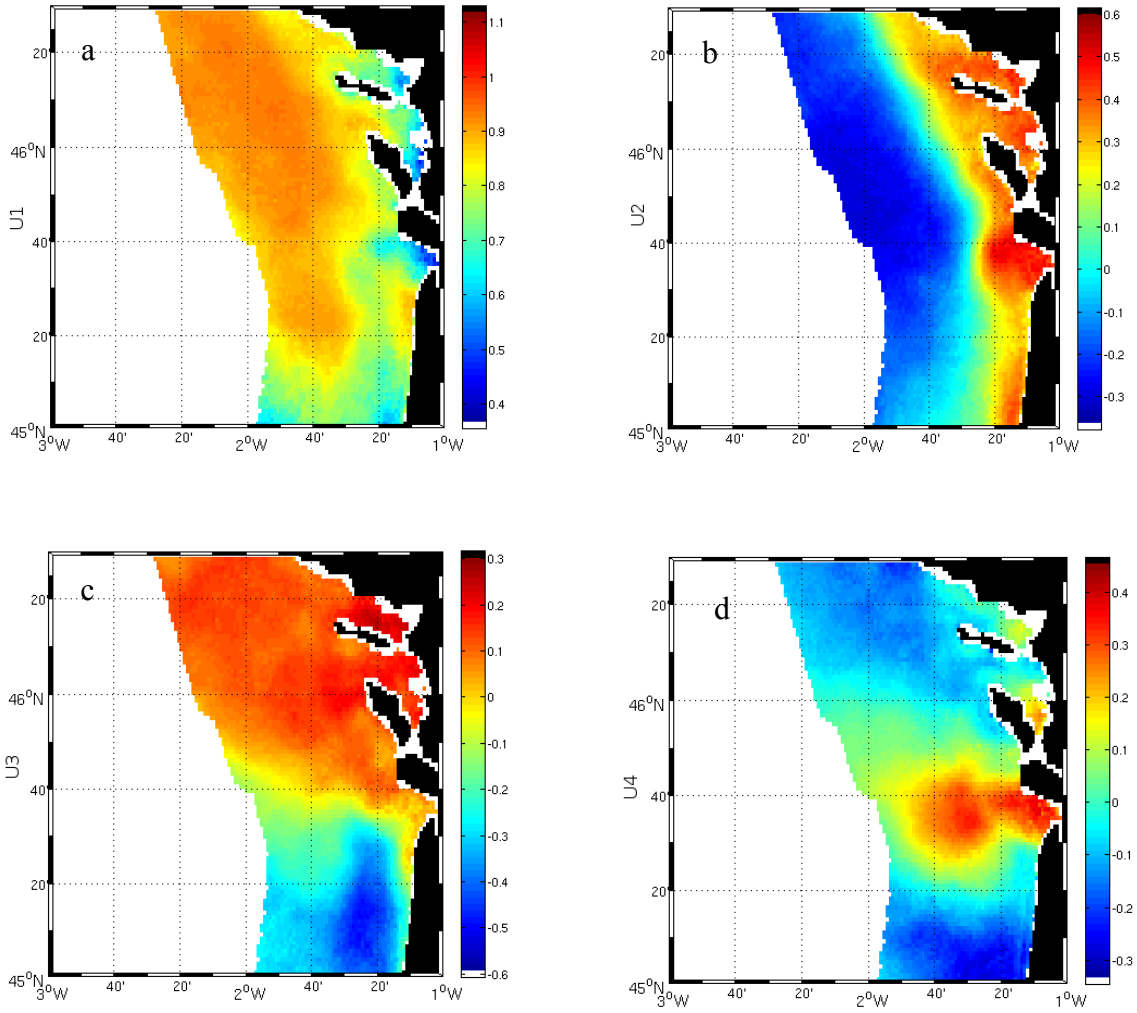
266 $$Cov(SPIM)=UVU^t \quad (17)$$

267 where U is a here 5682*5682 matrix containing the spatial modes (Eigenvectors) of the
268 covariance decomposition (ordered by percentage of explained variance). Associated with each
269 spatial mode k, its expansion coefficient (also referred in the literature as principal component) is
270 the time evolution of the k^{th} mode:

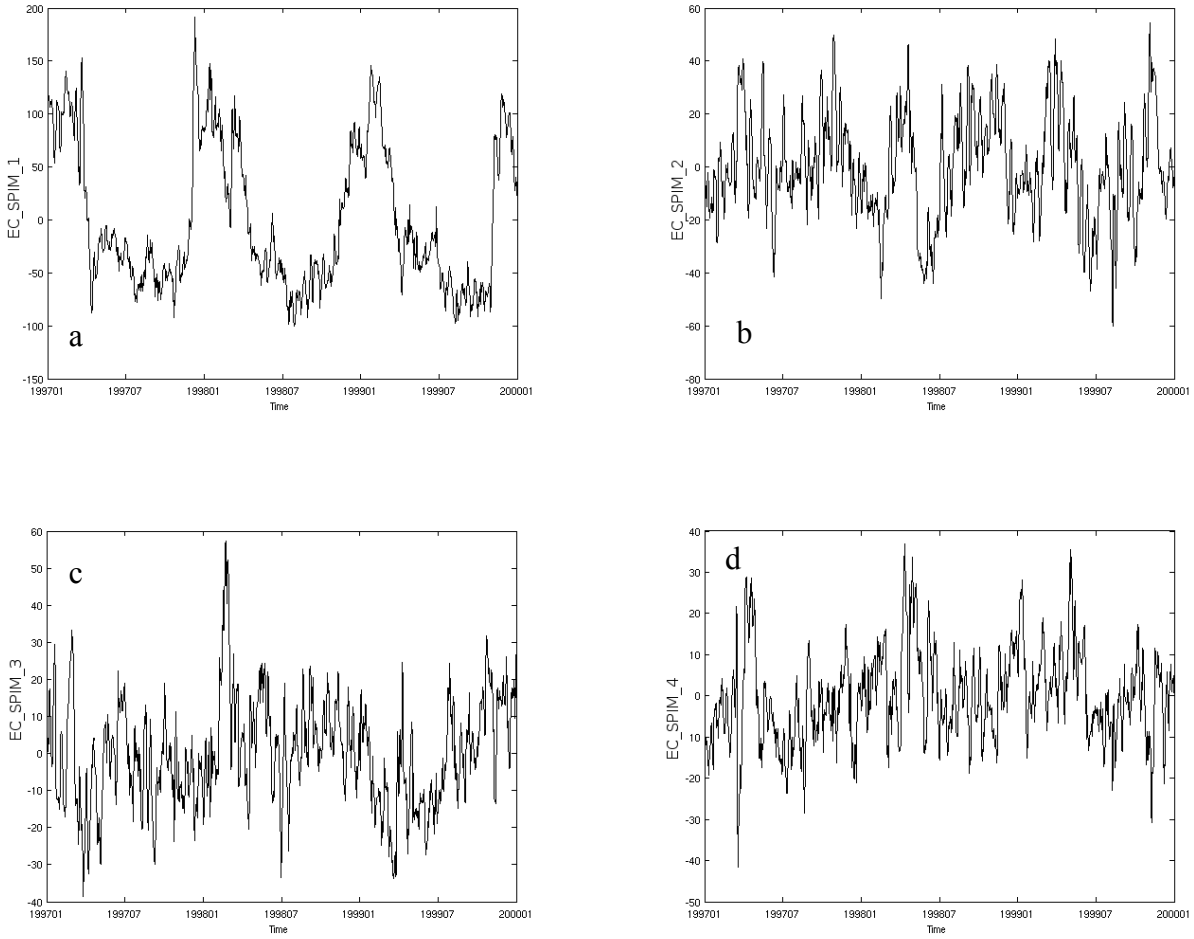
271 $$EC\_SPIM_{k,t} = SPIM_t * U_k \quad (18)$$

272 Figure 2 shows the four first spatial modes of the EOF decomposition. Figure 3 depicts the four
273 associated time series $EC\_SPIM_{i=1,4}$. The first mode (Fig.2a) comprises 85% of the total variance. It
274 clearly addresses the seasonal cycle as shown in Fig. 3a where the switch between winter (high
275 values of EC_SPIM_1 correspond here to high values of SPIM observed in winter) and summer
276 periods is clearly visible. The variability around the seasonal mean is captured by the other modes
277 (Fig.2 c-e & Fig 3 c-e). Mode 2 refers to the inter-annual and the intra-seasonal variability in the
278 shoreward gradient and represents 7% of the total variance. Mode 3 addresses some North-South

279 gradients and represents 4% of the total variance and mode 4 is clearly influenced by the Gironde
280 river (Fig. 2d), which brings sediments during water outflow, and represent 3% of the variance. By
281 construction, EOF decomposition imposes the orthogonality [21] of the spatial modes (Fig. 2).



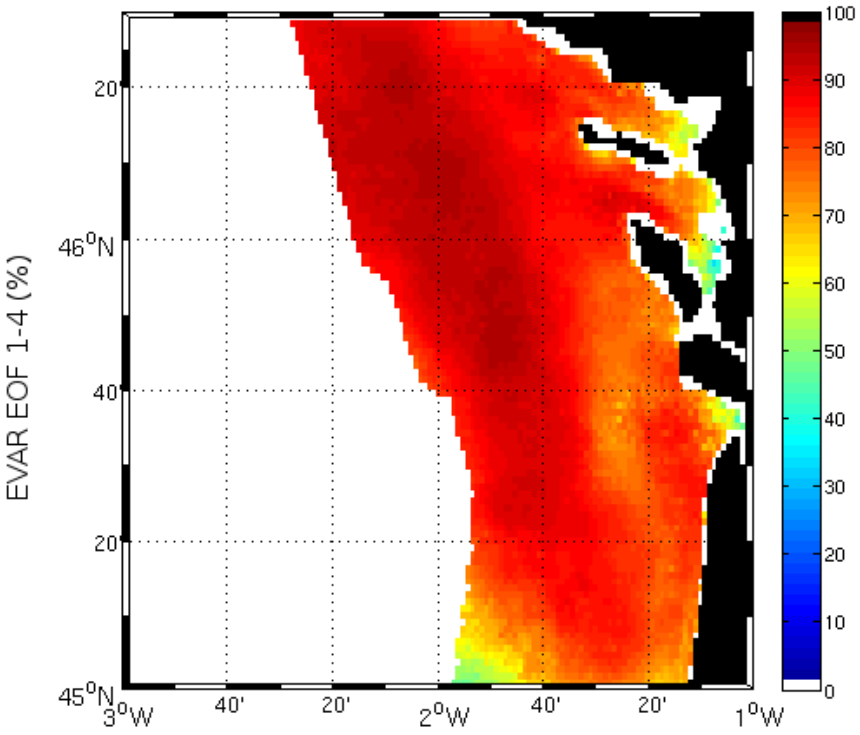
282 Figure 2: spatial modes of the EOF decomposition of the SPIM observed from satellite from
283 2007-2009 in the Gironde mouth river. From left to right and top to bottom the first four EOF
284 modes account respectively for 85, 7, 4 and 3% of the total variance.



285 Figure 3: EOF decomposition of the SPIM observed from satellite from 2007-2009 in the
286 Gironde mouth river: from left to right and top to bottom, the expansion coefficients ($EC\_SPIM_{1\text{-}4}$)
287 associated with the first four EOF modes depicted in Fig. 2, i.e. the time evolution of the spatial
288 modes. The reconstruction of the SPIM variable from the estimated ECs is performed as:

289 $$SPIM_t = k\text{EC\_SPIM} \,.\, Uk \qquad (19)$$

290 The total explained variance using the 4 first modes is shown Fig. 4. On average, the explained
291 variance represents 99 % of the total variability on the areas with some local minima of 60%
292 observed at the very near-shore and the Southwestern part of the area.



293

294 Figure 4: Variance explained by the four first modes of the EOF decomposition of the
295 suspended matters.

296

297 ## *4.2 Predictors and covariates*

298

299 The predictors X are the variables used in the estimation of Y and Z any time (Eq. 13 & 14). We
300 used here wave height (WH) daily means of the Wave Watch 3 model (WW3; [24, 36]) provided by
301 the IOWAGA and PREVIMER programs, Western and Northern winds interpolated from
302 QuickSCAT and ASCAT observations in conjunction with ECMWF forecasting [25], provided by
303 Ifremer, tide index (SHOM, 2000) at Bordeaux and the flow measurement of river la Gironde.
304 Similarly to the SPIM data, all the data were log transformed. For the wind data which is signed, the
305 transformed log variable was signed negatively a posteriori to the log transformation. The WH first
306 mode of the EOF decomposition explained 98 % of the total variance, 93% for the Northern wind

307 (WND1), and 96% for the Western wind (WND2). Covariates are the normalized predictors used in
308 the estimation but considered at t-2. This lag has been estimated as the optimal time-lag using BIC
309 and EVAR results on the training dataset.

310

311 # 5 Results

312

313 We summarize in Table 1 the prediction performance for the first four ECs of the SPIM issued
314 from four models: HMM, NHMM, HMM-AR, and NHMM-AR. The number of considered modes
315 for the mixture varies from 1 to 3. The one-mode models refer to a simple multivariate regression
316 analysis. For each configuration we provide the BIC and EVAR_train on the training dataset
317 (2007-2009) and EVAR_valid on the validation dataset (2010). Note that the selection of the
318 predictors and resulting covariates is achieved as a prior step as described in Section 2.6. The first
319 mode of the EOF decomposition explains 85% of the total variance. EC_WH_1 and EC_WND2_1
320 (respectively the expansion coefficient of the first EOF of the Western winds) are identified as
321 being the relevant predictors. This mode captures the mean seasonal variability of the SPIM, which
322 is mainly driven by WH and the North Atlantic storms and at a second order by the Western winds.
323 For EC_SPIM_1, when no autocorrelation term is used, the best fit is obtained for a 3-regime
324 NHMM model (BIC= 9873, EVAR_train=90% and EVAR_valid=85%). When a first order
325 autocorrelation term is added, the best fit is issued from a 3-regime HMM-AR model: BIC= 7997,
326 EVAR_train = 98% and EVAR_valid = 97%. The lag-1 autocorrelation value is 0.85 for
327 EC_SPIM_1, and therefore the weight given Y_{t-1} is important compared to the other covariates,
328 EC_WH_1 and EC_WND2_1. This stresses the fact that when available first autoregressive term
329 should be included to enhance the performances

pas clair

The second mode of the EOF decomposition of the SPIM variability explains 7% of the total variance. The selected predictors are the first mode of the Western wind, the tide, and the river flow. The variability captured by EC_SPIM$_2$ relates to the local Westward wind, which is not captured by the WH model, and the very coastal variability introduced by the tide and the river outflow. For the non-AR models the selected model was the three-regime NHMM. It is interesting to note in this case that EVAR_valid increased from 50 to 73% between the HMM and the NHMM, highlighting the contribution of the non-homogeneous transition model. By contrast, the HMM-AR performed slightly better than the NHMM-AR. niveau de contribution des variables complémentaires

Table 1: Model performance for each EOF mode of the SPIM variability. For each configuration we report the BIC (a) and the explained variance (EVAR_train, b) for the training dataset (2007-2009), and the explained variance (EVAR_valid, c) for the validation dataset (2010). In bold are highlighted for each EC the selected configurations (see § 5.2).

EC_SPIM	Number of modes, M					
	1	2	3	1	2	3
1	HMM (a) 11183 (b) 37 (c) 32	HMM 10037 84 70	HMM 9874 85 75	HMM-AR 8157 92 91	HMM-AR 7986 95 93	HMM-AR **7997** **98** **97**
	NHMM 11184 37 34	NHMM 10037 84 71	NHMM **9873** **90** **85**	NHMM-AR 8171 92 90	NHMM-AR 7994 92 94	NHMM-AR 8018 98 97
2	HMM 9403 18 12	HMM 8579 67 33	HMM 8129 76 50	HMM-AR 7167 90 87	HMM-AR 7098 91 89	HMM-AR **7075** **92** **91**
	NHMM 9451 18 12	NHMM 8614 67 44	NHMM **8152** **79** **73**	NHMM-AR 7188 89 88	NHMM-AR 7383 90 87	NHMM-AR 7070 92 90
3	HMM 8840 12 7	HMM 8222 57 44	HMM 7844 68 72	HMM-AR 6723 85 84	HMM-AR 6632 86 91	HMM-AR **6630** **88** **92**

	NHMM 8866 11 16	NHMM 8246 59 45	NHMM **7862** **75** **76**	NHMM-AR 6745 88 86	NHMM-AR 6673 88 91	NHMM-AR 6703 88 92
4	HMM 8248 18 28	HMM 7596 60 63	HMM 7285 71 72	HMM -AR 6398 85 86	HMM -AR 6416 85 86	HMM -AR **6313** **86** **86**
	NHMM 8276 18 28	NHMM 7628 62 59	NHMM **7267** **70** **75**	NHMM-AR 6426 85 83	NHMM-AR 6445 85 83	NHMM-AR 6357 86 85

342

343 The third mode of the EOF decomposition of the SPIM variability explains 4% of the total
344 variance. It captures some inter-annual and intra-seasonal variability of the latitudinal gradient of
345 the SPIM. The selected predictors are EC_WH_1, EC_WND1_1 (Northern) and the tide. Once again,
346 three-regime NHMM and HMM-AR provide the best results.

347 Regarding the fourth mode of the EOF decomposition of the SPIM variability, which accounts
348 3% of the total variance, EC_WH_1, EC_WND2_1, the tide and the river flow are selected as
349 contributive predictors. We reconstruct 75 % of EC_SPIM_3 variance of the validation dataset using
350 a three-regime NHMM and 86% using a three-regime HMM-AR. We note that globally, the three
351 indices (BIC, EVAR_train and EVAR_valid) tend to select the same models.

352

353 ## *5.1 Example with the estimation of EC_SPIM_1*

354

355 We report in Figure 5 the temporal evolution of the three regimes of the NHMM for EC_SPIM_1.
356 In table 2 are shown the corresponding coefficients for each predictor and the intercept. The first
357 regime (light grey), characterized by high SPIM levels (intercept of 65), is referred as a 'winter
358 regime'. The' winter regime' also strongly relates to the wave height (WH regression coefficient of
359 0.6). Dark grey periods (regime 3) are identified as a 'transition regime', and medium grey (regime

2) identified as the 'summer regime'. For regimes 2 and 3, the coefficients for WH decrease respectively to 0.12 and 0.09. In summer the energy brought by waves is not sufficient enough to re-suspend massively the sediments. It might be noticed that for all regimes the wind conditions show a small but significant effect on EC_SPIM_1. When an autocorrelation term is added (HMM-AR, table 2), the AR(1) coefficient value is 0.86 for the regime 1 (winter), and 0.9 for regime 2 and 3 which underlies the natural higher autocorrelation level of SPIM when the concentration is low.

Figure 4 compares the prediction of EC_SPIM_1 using a single multivariate regression (green) and the proposed multi-regime NHMM. In this case the explained variance value (Table 1) is of 37% for the multivariate regression model compared to 85% for the three-regime NHMM.

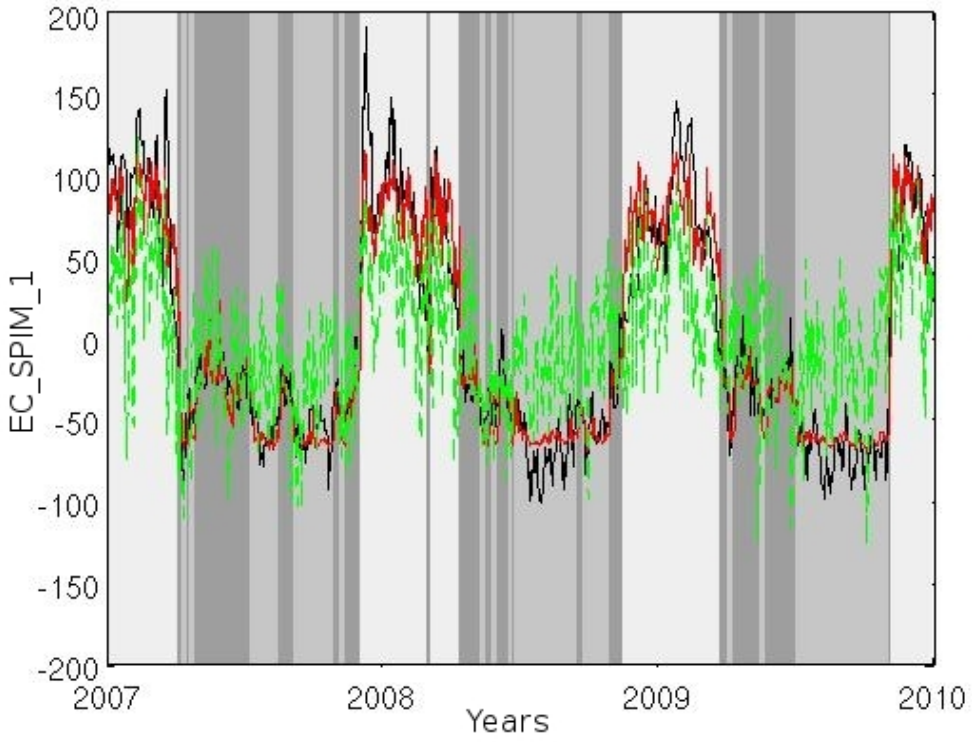


Figure 5: Estimation of the EC_SPIM_1 (in black) using EC_WH_1, EC_WND2_1 and a single regression (green) and a 3 regime NHMM (red). The nuances of grey in the background highlight the temporal distribution of the regimes.

Table 2: Estimated regression parameters for each of the three regimes of the NHMM and the HMM-AR for the first mode of the SPIM EOF decomposition: regression parameters involve an intercept and the regression coefficients of the wave height and western wind velocity.

	EC_WH_1	EC_WND_2	Intercept	
NHMM	(1, winter) 0.6037 (2, summer) 0.0910 (3, transition) 0.1210	-0.0632 0.0006 0.0100	65.0672 -61.6442 -24.6578	
	EC_WH_1	EC_WND_2	Intercept	AR(1)
HMM-AR	(1) 0.2383 (2) -0.0050 (3) 0.0354	-0.0033 0.0168 0.0035	4.4694 -3.9531 0.6079	0.86 0.90 0.90

variance expliqué par chaque terme

Figure 6 illustrates the non-homogeneous transition used in the NHMM between the 'transition' (Z_t=3) and 'winter' (Z_t=1) regimes. The probability of switching from regime 3 to 1 increases with wave height covariate WH_1 with a probability of switching close to zero when WH_1 is negative and a probability close to one for large WH_1 values.

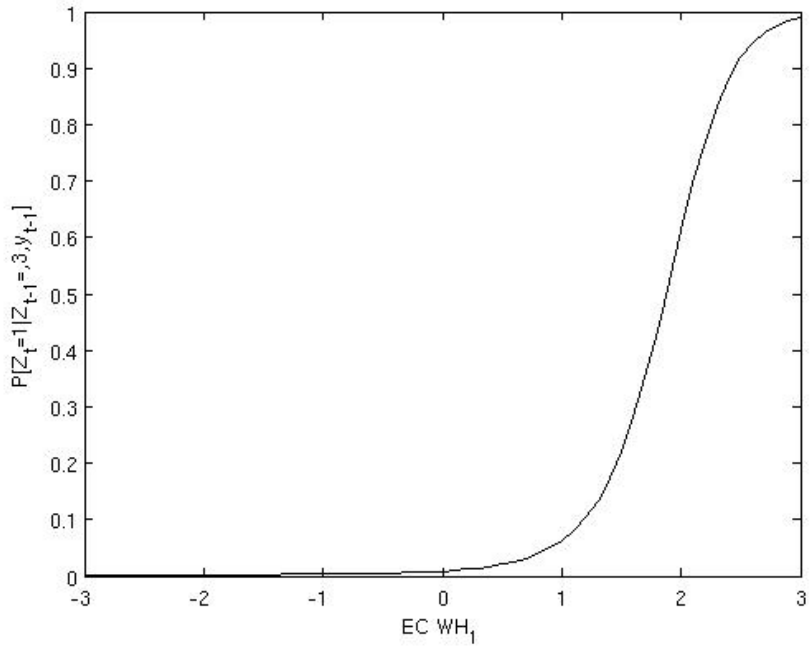
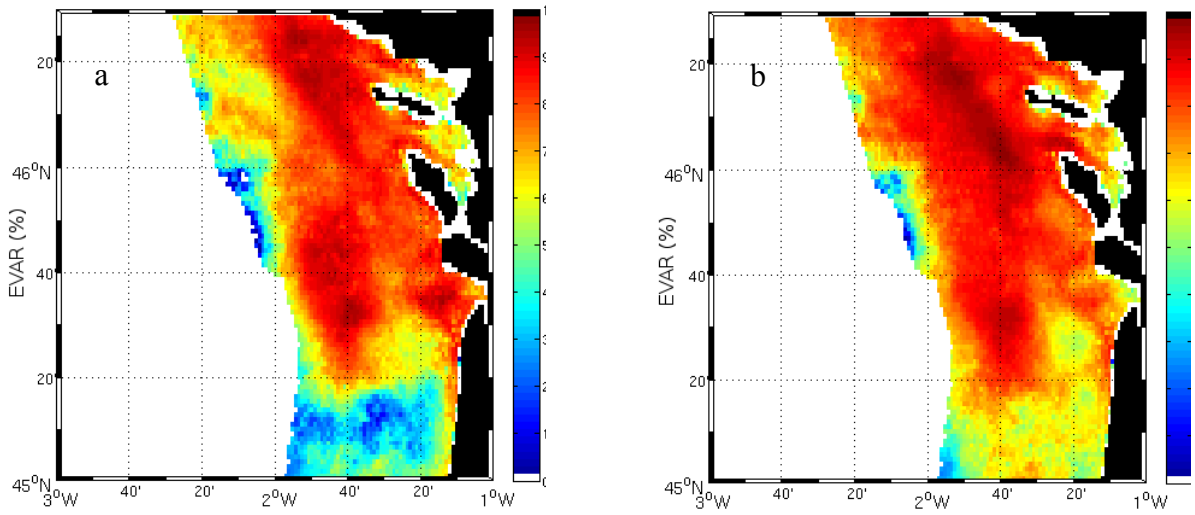


Figure 6: Non-homogeneous transition between 'transition regime' (Zt=3) (light grey Fig. 5) and 'winter regime' (Zt=1) (light grey Fig. 5) as a function of the wave height covariate WH_1.

5.2 Forcasting of the SPIM on the 2010 validation dataset

We forecast SPIM fields from the reconstructed ECs and the selected models (Table 1 & Eq. 19). Figure 7a&7b compare EVAR_valid of the initial field (SPIM) using the three-regime NHMM and NHMM-AR models selected in Table 1 for their results. On average we were able to predict at t+1 80% of the variance using the NHMM (Fig. 7a) and 90% using the NHMM-AR. The spatial distribution of the error is not homogeneous. Fig. 7a shows that EVAR_valid value is of 90% in the Northern part with nevertheless poorer results in the South. Fig. 7b shows that the AR_1 component of the model increases EVAR for the whole area.

We also considsr the results of a standard multi-regression analysis. If only one regime is considerer NHMM and HMM resort to a strandard multivariate regression and NHMM-AR and HMM-AR to a strandard multivariate regression including an AR_1 coefficient the transition probabilities being equal to 1. Fig. 7c shows the obtained results with the standard multivariate regression and Fig 7d the standard multivariate regression including an AR_1. From Fig. 7c to 7a, the gain in explained variance is in mean about 250% (from in mean 32% Fig. 7c to 80% Fig. 7a) while for the AR models, the gain is about from 11% (from in mean 83% Fig. 7d to 93% Fig. 7b).



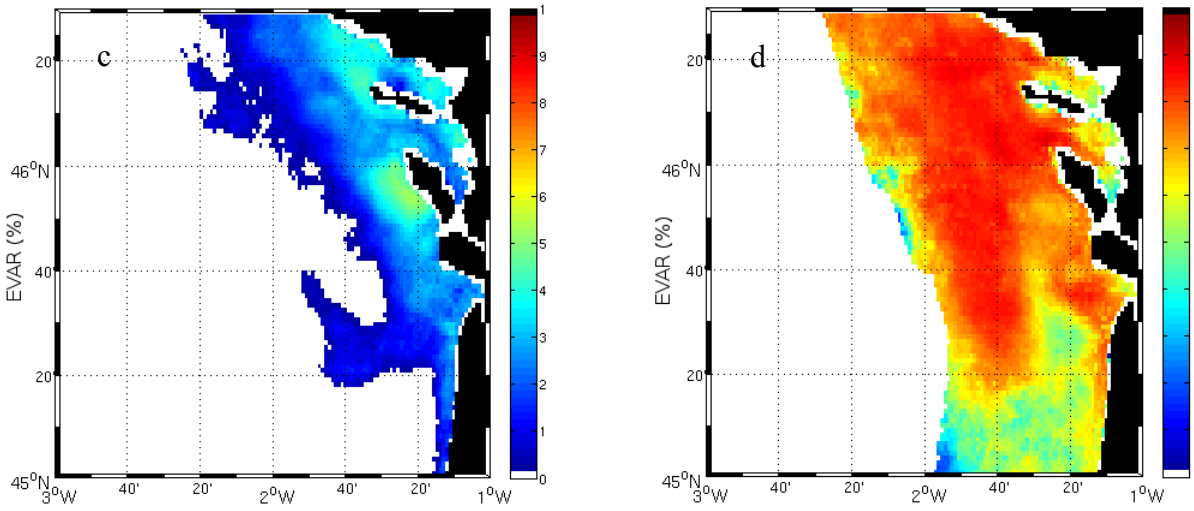


Figure 7: Explained variance for the 2010 validation dataset reconstructed using the selected 3-regime NHMM (a) and NHMM-AR (b), compared with the standard multivariate regression without AR_1 (c) and including an AR_1 (d).

To consider the model forecast performances, we report the short-term forecast results at different time steps (cf. Eq.13 & 14). For the HMM-AR and NHMM-AR (Eq. 14) is the estimated value, the observation being not available (see § 3.2). Table 3 synthesizes the explained variance statistics using 3 regimes and the four tested models for the forecasting at t+1, t+5 and t+15 and the validation 2010 dataset (Eq. 19).

The long term forecasting results are globally better with the NHMM. At t+15 using the NHMM we are able to forecast 74% of the variance for 2010, compared with 40% for the HMM. In this case where the covariates and predictors (mode outputs for which the short term predictions are assumed to be available) are used in the estimation of the regime transition probability . For autoregressive models, at t+15, we were able to forecast 75% of the 2010 variance with the NHMM-AR compared with 70% with the HMM-AR. For the NHMM-AR the covariates help in the estimation of both and . At t+15 NHMM and NHMM-AR show equivalent results underlying the maximal time-step to consider for the between these two models.

Table 3: Validation results on year 2010. Explained variance (Eq. 16) for the forecast at t+1, t+5 and t+15 of the 2010 validation dataset.

EVAR for the 2010 validation dataset	
HMM (t+1) 73 (t+5) 63 (t+15) 40	HMM-AR 93 80 70
NHMM (t+1) 80 (t+5) 77 (t+15) 74	NHMM-AR 93 82 75

SVR model was also evaluated to evaluate the performances of a non-linear model on the studied dataset. To perform the comparison, we train the SVR model (http://www.csie.ntu.edu.tw) for each EC using the same training dataset (2007-2009) and performed forecasting using the same validation dataset (2010). We used the setting as following: model epsilon-SVR (s=3), linear or polynomial kernel (t=0 or 1) and the same inputs (predictors, covariables) for each EC. Parameters c and g [10] were optimised for each EC using the training dataset and the cross validation mode. On the 2010 validation dataset, better forecast results reached 40% at t+1 of the EVAR (without AR) and 85% with an AR coefficient. The results were significantly worse than those obtained using the time-varying models for increasing time steps. The SVR can address non-linear relationships, it cannot nevertheless deal with multi-regime processes. By contrast, the latent-regime model addresses by nature multi-regime processes and can approximate non-linear relationships as a series of linear models.

gaussien?

6 Discussion

We investigated the relevance of four regime-switching latent regression models, namely HMM, NHMM, HMM-AR and NHMM-AR to characterize time-varying linear relationships between the high resolution inorganic suspended matter concentration (SPIM), estimated from 2007 to 2009 using MODIS, SeaWiFS and MERIS data in a coastal area, and its forcing conditions, i.e. the wave height, the Northern and Western winds, the tides and the river flow. The estimated regimes are then used to forecast the SPIM using the independent year 2010 dataset, from t+1 to t+15.

An optimal number of three distinct geophysical regimes were needed to capture the different dynamics and optimize forecasting performance. Autoregressive and non-homogeneous model showed better performances. With the evaluated models and the 2010 validation dataset we were able to forecast at t+1 80 % of the variance explained using a NHMM and 93 % using a NHMM-AR. In the latest case the strong natural autocorrelation of the studied signal is an important predictor to consider. The explained variance on the prediction at +1 for a standard multivariate regression was of 32% and 80% (with an AR_1 term). Using a SVR we were able to forecast at t+1 respectively 40% and 85% (with an AR_1) of the explained variance.

As illustrated for the first SPIM EOF component (Figure 5), the proposed multi-regime setting allowed us to identify three seasonally varying relationships between the observed turbidity, the wave height and the wind. We did not drive the model to account for seasonal regimes but we identified three seasonally-discriminated regimes, with two leading factors: the mean SPIM level (intercept) and the Western wave height. These regimes identified directly physical behaviors, here the minimum of energy to be brought by the Western swell to re-suspend the sediments. This is regarded as a key feature of the latent-regime model compared to other non-linear regression models, such as Neural Networks [38] or SVR [10] which cannot address multi-regime

relationships and are hardly interpreted in general. Using our dataset the non-linear SVR was not able to retrieve the regime changes.

Regarding long-term forecast performance, at t+15 best results obtained were of 74% of explained variance for the NHMM and 75% for the NHMM-AR. For short period, typically from 1 to 15 days, when the observed Y is not available, NHMM-AR provided the best results. In this case the predictors and covariates are used in the estimation of both and . At t+15 NHMM and NHMM-AR showed similar results underlying the maximal time-step to consider, when no observation of Y is available, for the choice between these two models.

In the future, we will address the forecasting of the chlorophyll-a using satellite-derived observations such as the photosynthetic available radiation, the temperature, the suspended matters (as index of available nutrients) and light attenuation [39]. In this more complicated case, second order relationships between the variable and its predictors have to be evaluated, the chlorophyll-a dynamic being not anymore a passive result of the forcing conditions, as expected with the SPIM, but having its proper characteristics depending on each phytoplankton specie. Extensions of the considered latent regime setting to other inverse problems in satellite sensing data analysis are also under investigation, such as latent regime inversion procedures for satellite-derived chlorophyll-a concentration to account for different water types (turbid or not turbid) and/or the presence of specific phytoplankton species.

7 Acknowledgements

The authors thank Aldo Sottolichio from the Université of Bordeaux 1, for the provision of the in-situ Gironde flow measurement, Pierre Tandeo for fruitful advises and the MCGS (Marine Collaborative Ground Segment; http://www.mcgs.fr) project which aim at making the most of ESA Sentinels satellites potential for users driven services based on high level products. MCGS

addresses the need of the European Space Agency to build up data processing centers in conjunction with the Copernicus Program for the provision of services to local and national, public and private European institutions or entities involved in marine activities. The project is co-funded by the French Government (Fonds Unique Inter-ministériel), local authorities of the Bretagne and Provence-Alpes-Côte d'Azur regions and the European Regional Development Fund (ERDF), under support of the French Space Agency (CNES).

8 References

1 P. Lazure, V. Garnier, F. Dumas, C. Herry, M. Chifflet. "Development of a hydrodynamic model of the Bay of Biscay". Validation of hydrology. Continental Shelf Research, 29(8), 985-997, 2009.

2 L. Debreu, P. Marchesiello, P. Penven and G. Cambon, "Two-way nesting in split-explicit ocean models: algorithms, implementation and validation", Ocean Model., 49-50, 1-21, 2012.

3 A. Sottolichio, P. Le Hir, P. Castaing. "Modeling mechanisms for the turbidity maximum stability in the Gironde estuary, France", Coastal and Estuarine Fine Sediment Processes, pp 373-386, 2000.

4 A. Rivier, F. Gohin, P. Bryere, C. Petus, N. Guillou, G. Chapalain. " Observed vs. predicted variability in non-algal suspended particulate matter concentration in the English Channel in relation to tides and waves". Geomarine Letters, 32(2), 139-151, 2012.

5 C. Aitken, "On Least Squares and Linear Combinations of Observations", Proceedings of the Royal Society of Edinburgh, 55, 42–48, 1935.

6 William E. Wecker, Craig F. Ansley, The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing Journal of The American Statistical Association , 78(381):81-89, 1983.

7 Box, George; Jenkins, Gwilym, Time Series Analysis: forecasting and control, rev. ed., Oakland, California: Holden-Day, 1976.

8 Tesfaye, Y. G., Meerschaert, M. M., and Anderson, P. L. Identification of periodic autoregressive moving average models and their application to the modeling of river flows. Water Resources Research, 42(1), 2006

9 Some Neural Network applications in environmental sciences part I: forward and inverse problems in geophysical remote measurements, V. Krasnopolsky and H. Schiller, Neural Networks, vol 16 (2003), 321-334.

10 Chih-Chung C., Chih-Jen L. LIBSVM: A library for support vector machines Transactions on Intelligent Systems and Technology (TIST), 2011.

11 P. Ailliot, V. Monbet. "Markov-switching autoregressive models for wind time series". Environmental Modelling & Software, 30, 92-101, 2012.

12 F. Gohin, S. Loyer, M. Lunven, C. Labry, J.M. Froidefond, D . Delmas, M . Huret, A. Herbland, "Satellite-derived parameters for biological modelling in coastal waters: illustration over the eastern continental shelf of the Bay of Biscay". Remote Sensing of Environnement 95(1), 29–46, 2005.

13 D. Doxaran, « Télédétection et modélisation numérique des flux sédimentaires dans l'estuaire de la Gironde », PhD thesis, University Bordeaux 1, France, 326 pp, 2002.

14 D. Doxaran, J.M. Froidefond, P. Castaing and M. Babin. "Dynamics of the turbidity maximum zone in a macrotidal estuary (the Gironde, France): Observations from field and MODIS satellite data", Estuarine Coastal and Shelf Science, 81, 321–332, 2009.

15 D.G. Bowers, C.E. Binding."The optical properties of mineral suspended particles: a review and synthesis". Estuarine Coastal Shelf Science, 67(1/2), 219–230, 2006.

16 D. Eisma, P. Bernhard, G.C. Cadée, V. Ittekkot, J. Kalf, R. Laane, J.M. Martin, W.G. Mook, A. Van Put, T. Schuhmacher, "Suspended matter particle size in some West-European estuaries; part II: a review on floc formation and break up", Netherlands Journal of Sea Research, 28(3), 215–220, 1991.

17 W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear 469 regression", Journal of Classification, 5, 249–282, 1998

531 18 P. Tandeo, B. Chapron, S. Ba, E. Autret, R. Fablet, “Segmentation of Mesoscale Ocean Surface
532 Dynamics Using Satellite SST and SSH Observations Geoscience and Remote Sensing”, IEEE Transactions
533 on Volume, pp 1 – 9, 2013.

534 19 B.H. Juang, and L.R. Rabiner. “Hidden Markov models for speech recognition”, Technometrics, 33,
535 251-272, 1991.

536 20 C. Frankignoul and K. Hasselmann, “Stochastic climate models. Part II: Application to SST
537 anomalies and thermocline variability”. Tellus, 29, 289-305, 1977.

538 21 R.W. Preisendorfer, “Principal Component Analysis in Meteorology and Oceanography”, Elsevier,
539 New York,. pp 425, 1988.

540 22 Cressie NAC, Wikle CK. Statistics for Spatio-Temporal Data. Wiley; New York: 2011.

541 23 Abdi, H. "Discriminant correspondence analysis." In: N.J. Salkind (Ed.): Encyclopedia of
542 Measurement and Statistic. Thousand Oaks (CA): Sage. pp. 270–275, 2007.

543 24 F. Ardhuin, E. Rogers, A.V. Babanin, J. Filipot, R. Magne, A. Roland, A. van der Westhuysen, P.
544 Queffeulou, J.M. Lefevre, L. Aouf, F. Collard, “Semi-empirical dissipation source functions for ocean
545 waves. Part I: Definition, calibration, and validation”. Journal of Physical Oceanography, 40(9):1917–1941,
546 2010.

547 25 A. Bentamy, D. Croizé. Fillon, “Gridded Surface Wind Fields from
548 Metop/ASCAT Measurements”, International Journal of Remote Sensing, 33:
549 1729-1754, 2011.

550 26 Courants de marée et hauteurs d’eau. La Manche de Dunkerque à Brest. Service Hydrographique et
551 Océanographique de la Marine, Brest, Rapport 564-UJA, 2000

552 27 Tong, H. Non-linear time series, a dynamical systems approach. Oxford University Press, 1990.

553 28 J. P. Hughes and P. Guttorp. “A Class of Stochastic Models for Relating Synoptic Atmospheric
554 Patterns to Regional Hydrologic Phenomena”, Water Resources Research, 30, 1535-1546, 1994a.

555 29 J. P. Hughes and P. Guttorp, “Incorporating Spatial Dependence and Atmospheric Data in a Model
556 of Precipitation”. Journal Applied Meteorology, 33, 1503-1515, 1994b.

557 30 A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the EM
558 algorithm”, Journal of the Royal Statistical Society. Series B (Methodological), 39, 1–38, 1977.

559 31 F. Castino, R. Festa, and C.F. Ratto, “Stochastic modelling of wind velocities time series”, Journal of
560 Wind

561 32 W. Zucchini and P. Guttorp. “A hidden Markov model for space-time precipitation”, Water
562 Resources Research, 27, 1917–1923, 1991.

563 33 O. Cappe, E. Moulines, and T. Ryden. “Inference in hidden Markov models”. Springer-Verlag, New
564 York, 2005.

565 34 H.S. Bhat, N. Kumar, “On the derivation of the Bayesian Information Criterion”, 2010, Available
566 from http://nscs00.ucmerced.edu/~nkumar4/BhatKumarBIC.pdf.

567 35 B. Saulquin, F. Gohin, R Garrello, “Regional objective analysis for merging high-resolution MERIS,
568 MODIS/Aqua, and SeaWiFS chlorophyll-a data from 1998 to 2008 on the European Atlantic shelf”, IEEE
569 Trans Geoscience Remote Sensing 49(1), 143–154, 2011.

570 36 H.L. Tolman. “A mosaic approach to wind wave modeling”. Ocean Modell, 25(1/2), 35–47, 2008.

571 37 B. Saulquin, R. Fablet, A. Mangin, G. Mercier, D. Antoine, and O. Fanton d'Andon, “Detection of
572 linear trends in multisensor time series in the presence of autocorrelated noise: Application to the
573 chlorophyll-a SeaWiFS and MERIS data sets and extrapolation to the incoming Sentinel 3-OLCI mission”,
574 Journal of Geophysical Research Oceans, 118, 3752–3763, 2013.

575 38 H. Schiller, R. Doerffer, “Neural network for emulation of an inverse model operational derivation
576 of Case II water properties from MERIS data”. International Journal of Remote Sensing, 20(9), 1735-1746,
577 1999.

578 39 B. Saulquin, A. Hamdi, F. Gohin, J. Populus, A. Mangin, O. Fanton D'Andon, “Estimation of the
579 diffuse attenuation coefficient K-dPAR using MERIS and application to seabed habitat mapping”. Remote
580 Sensing Of Environment, 128, 224-40

581